Trustworthy AI: From Principles to Practices

Year: 2021 | Citations: 493 | Authors: Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu

---

Abstract

The rapid development of Artificial Intelligence (AI) technology has enabled the deployment of various systems based on it. However, many current AI systems are found vulnerable to imperceptible attacks, biased against underrepresented groups, lacking in user privacy protection. These shortcomings degrade user experience and erode people's trust in all AI systems. In this review, we provide AI practitioners with a comprehensive guide for building trustworthy AI systems. We first introduce the theoretical framework of important aspects of AI trustworthiness, including robustness, generalization, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability. To unify currently available but fragmented approaches toward trustworthy AI, we organize them in a systematic approach that considers the entire lifecycle of AI systems, ranging from data acquisition to model development, to system development and deployment, finally to continuous monitoring and governance. In this framework, we offer concrete action items for practitioners and societal stakeholders (e.g., researchers, engineers, and regulators) to improve AI trustworthiness. Finally, we identify key opportunities and challenges for the future development of trustworthy AI systems, where we identify the need for a paradigm shift toward comprehensively trustworthy AI systems.